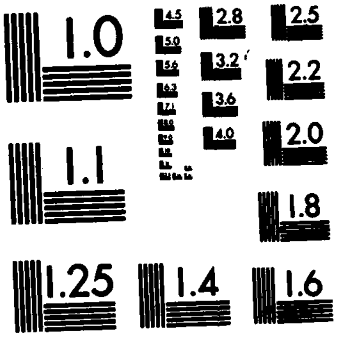
AD-A125 155 ENTROPY INTERPRETATION OF GOODNESS OF FIT TESTS(U) 1/1
TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS
E PARZEN JAN 83 TR-B-8 ARO-16992.14-MA DAAG29-80-C-0070

UNCLASSIFIED F/G 12/1 NL

END
FILMED
DTIC


1.0
4.5
5.0
5.6
6.3
7.1
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

TEXAS A&M UNIVERSITY

COLLEGE STATION, TEXAS 77843-3143

INSTITUTE OF STATISTICS
Phone 713 - 845-3141

ENTROPY INTERPRETATION OF GOODNESS OF FIT TESTS

by Emanuel Parzen

Institute of Statistics

Texas A&M University

Technical Report No. B-8

January 1983

Texas A&M Research Foundation
Project No. 4226

"Robust Statistical Data Analysis and Modeling"

Sponsored by the U. S. Army Research Office
Grant DAAG29-80-C-0070

Professor Emanuel Parzen, Principal Investigator

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
B-8

2. GOVT ACCESSION NO.
AD-A125155

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
ENTROPY INTERPRETATION OF GOODNESS OF FIT TESTS

5. TYPE OF REPORT & PERIOD COVERED
Technical

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Emanuel Parzen

8. CONTRACT OR GRANT NUMBER(s)
DAAG 29-80-C-0070

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Texas A&M University
Institute of Statistics
College Station, TX 77843

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

12. REPORT DATE
January 1983

13. NUMBER OF PAGES
13

14. MONITORING AGENCY NAME & ADDRESS(If different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Entropy-based statistical inference, goodness of fit tests, test for normality, Shapiro-Wilk statistic, quantile, density-quantile, quantile-density, autoregressive density estimator.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper describes a synthesis of statistical reasoning called FUN.STAT (because it is fun; functional (useful); based on functional analysis; estimates functions; and all graphs are of functions). UFN.STAT has three important components: quantile and density-quantile signatures of populations, entropy and information measures, and functional statistical inference. A FUN.STAT approach to the problem of identifying the probability distribution F(x) of a random variable X from a random sample is outlined.

ENTROPY INTERPRETATION OF GOODNESS OF FIT TESTS

Emanuel Parzen
Institute of Statistics
Texas A&M University

ABSTRACT. This paper describes a synthesis of statistical reasoning called FUN.STAT (because it is fun; functional (useful); based on functional analysis; estimates functions; and all graphs are of functions). FUN.STAT has three important components: quantile and density-quantile signatures of populations, entropy and information measures, and functional statistical inference.

A FUN.STAT approach to the problem of identifying the probability distribution F(x) of a random variable X from a random sample is outlined. To identify F_0 in the location-scale parameter model $F(x) = F_0((x-\mu)/\sigma)$, we estimate entropy difference $\Delta = H^0(f) - H(f)$. H(f) is Shannon entropy and $H^0(f) = \log \sigma + H(f_0)$ is entropy of the assumed model (which may maximize entropy). Estimators $\hat{H}_1$, $\hat{H}_2$, $\hat{H}_3$ of H(f) are defined which are respectively fully parametric, fully non-parametric, and parametric-select. Significance levels for $\hat{\Delta}$ are obtained by Monte Carlo methods. The family of parametric-select estimators of Δ may provide optimum tests of F_0 (such as normal or exponential) and estimators of F when one rejects F_0.

KEY WORDS: Entropy-based statistical inference, goodness of fit tests, test for normality, Shapiro-Wilk statistic, quantile, density-quantile, quantile-density, autoregressive density estimator.

1. INTRODUCTION.

Let $X_1,\ldots,X_n$ be a random sample of a continuous random variable X with distribution function $F(x) = \Pr[X \le x]$, $-\infty<x<\infty$, and quantile function $Q(u) = F^{-1}(u)$, $0 \le u \le 1$. Tests of normality or exponentiality are special cases of a location-scale parameter model, which we denote by the hypothesis

$$H_0\colon F(x) = F_0(\tfrac{x-\mu}{\sigma}),\ Q(u) = \mu + \sigma\, Q_0(u)$$

where $F_0(x)$ is a specified distribution with quantile function $Q_0(u)$. Table 1 lists F_0 and Q_0 for various standard distributions.

Research supported by the U. S. Army Research Office Grant DAAG 29-80-C-0070.

Table 1. STANDARD DISTRIBUTION FUNCTIONS AND QUANTILE FUNCTIONS

Name	$F_0(x)$	$Q_0(u)$
Normal	$\Phi(x) = \int_{-\infty}^{x} \phi(y)\, dy$; $\phi(x) = (2\pi)^{-\frac{1}{2}} \exp - \frac{1}{2} x^2$	$\Phi^{-1}(u)$
Exponential	$1 - e^{-x}$	$\log (1-u)^{-1}$
Weibull, Quantile shape parameter β	$1 - e^{-x^c}$, $x \geq 0$ $c = \frac{1}{\beta}$	$\left\{\log (1-u)^{-1}\right\}^{\beta}$
Extreme value of minimum	$1 - e^{-e^x}$ $-\infty<x<\infty$	$\log \log (1-u)^{-1}$
Extreme value of maximum	$e^{-e^{-x}}$ $-\infty<x<\infty$	$- \log \log u^{-1}$
Log normal	$\Phi(\log x)$, $x>0$	$\exp \Phi^{-1}(u)$
Logistic	$1-(1+e^x)^{-1}$	$\log \frac{u}{1-u}$

Many statistics have been introduced by statisticians to test the composite (location and scale parameters unspecified) hypothesis of normality. A superior omnibus test of normality (in terms of power) seems to be provided by a test statistic $W = \hat{\sigma}_2/\hat{\sigma}_1$, where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are scale estimators defined as follows: $\hat{\sigma}_1$ is sample standard deviation, while $\hat{\sigma}_2$ is a linear combination of order statistics estimator of σ. We call W a statistic of Shapiro-Wilk type because it is a variant of a test introduced by Shapiro and Wilk (1965) and Shapiro and Francia (1972).

The question arises: to discover a motivation for the W statistic which explains the source of its power, and to use this insight to extend W to other distributions F_0. In this paper we propose that the power of W can be explained by representing it as an "entropy difference" test statistic. We show that the test statistic for normality introduced by Vasicek (1977) is also an entropy difference statistic, as are test statistics introduced in Parzen (1979).

2. INFORMATION DIVERGENCE AND ENTROPY.

To compare two distribution functions F(x) and G(x) with probability densities f(x) and g(x), a useful measure is *information divergence*, defined by

$$I(f;g) = \int_{-\infty}^{\infty} \{-\log \frac{g(x)}{f(x)}\} \, f(x) \, dx$$

It can be decomposed into *cross-entropy*

$$H(f;g) = \int_{-\infty}^{\infty} \{-\log g(x)\} \, f(x) \, dx$$

and *entropy*

$$H(f) = H(f;f) = \int_{-\infty}^{\infty} \{-\log f(x)\} \, f(x) \, dx$$

by the important identity

$$0 \leq I(f;g) = H(f;g) - H(f).$$

To estimate entropy it is useful to express it in terms of the *quantile density* function q(u) and *density-quantile* function fQ(u) defined by

$$q(u) = Q'(u), \quad fQ(u) = f(Q(u)) = \{q(u)\}^{-1}$$

By making the change of variable u = F(x) one can show that

$$H(f) = \int_0^1 - \log fQ(u) \, du$$

$$= \int_0^1 \log q(u) \, du.$$

Under the hypothesis H_0 that $F(x) = F_0((x-\mu)/\sigma)$, a location-scale model, $q(u) = \sigma q_0(u)$ and

$$H(f) = \log \sigma + H(f_0).$$

3. ENTROPY DIFFERENCE TO TEST GOODNESS OF FIT. To test the hypothesis H_0 we propose to investigate (and eventually establish how to use optimally) test statistics which are entropy-difference statistics

$$\Delta(f) = H^0(f) - H(f)$$

where $H^0(f)$ is a parametric evaluation of the entropy of f, evaluated under the assumption that it obeys H_0, defined by

$$H^0(f) = \log \sigma + H(f_0),$$

while H(f) is a non-parametric evaluation of H(f), usually most conveniently obtained by

$$H(f) = \int_0^1 \log q(u)\, du \quad .$$

To estimate H(f) we have three types of estimators which we call

$\hat{H}_1$ fully parametric estimator,

$\hat{H}_2$ fully non-parametric estimator,

$\hat{H}_3$ smooth or parametric select estimator

Similarly to estimate $H^0(f)$ we have several types of estimators depending on the estimator $\hat{\sigma}_j$ we adopt for σ; thus

$$\hat{H}^0{}_j = \log \hat{\sigma}_j + H(f_0)$$

Three important possibilities for $\hat{\sigma}_j$ are:

$\hat{\sigma}_1$ maximum likelihood estimator,

$\hat{\sigma}_2$ optimal linear combination of order statistics estimator

$\hat{\sigma}_3$ estimator of score deviation $\sigma_3 = \int_0^1 f_0Q_0(u)\, q(u)\, du$.

Under H_0 these estimators are all asymptotically efficient estimators of σ.

While one can conceive of about 9 possible estimators of the entropy difference Δ, we discuss only three estimators which we denote $\hat{\Delta}_{11}$, $\hat{\Delta}_{12}$, :nd $\hat{\Delta}_{33}$.

4. ENTROPY-DIFFERENCE INTERPRETATION OF SHAPIRO-WILK STATISTIC

To test the hypothesis H_0: X is $N(\mu,\sigma^2)$, a test statistic W of Shapiro-Wilk type is of the form

$$W = \hat{\sigma}_2 \div \hat{\sigma}_1$$

where $\acute{\sigma}_1$ is the sample standard deviation and

$$\hat{\sigma}_2 = \sum_{j=1}^{n} \Phi^{-1}\left(\frac{j-0.5}{n}\right) X_{(j)} \div \left\{\sum_{j=1}^{n} \left| \Phi^{-1}\left(\frac{j-0.5}{n}\right)\right|^2\right\}^{\frac{1}{2}}$$

is an asymptotically efficient estimator of σ based on linear combinations of the order statistics $X_{(1)} < \ldots < X_{(n)}$ of the random sample. The first step in the entropy interpretation of W is to consider instead the statistic

$$\hat{\Delta}_{11} = - \log W = \log \hat{\sigma}_1 - \log \hat{\sigma}_2 = \hat{H}_1^0 - H_1$$

where [with $f_0(x) = \phi(x) = (2\pi)^{-\frac{1}{2}} \exp -(\frac{1}{2})\, x^2$, and $H(f_0) = \frac{1}{2}(1 + \log 2\pi)$]

$$\hat{H}_1^0 = \log \hat{\sigma}_1 + H(f_0)$$

is an estimator of $H^0(f)$ based on $\hat{\sigma}_1$, and $\hat{H}_1$ is a purely parametric estimator of $H(f)$ based on the parametric estimator $\hat{\sigma}_2$; note $\hat{H}_1 = \hat{H}_2^0$.

Significance levels for the entropy-difference statistic $\hat{\Delta}_{11} = - \log W$ are obtainable from tables of the W statistic [for example, Filliben (1975)]. An example of 5% significance levels (for accepting normality) are

$$\hat{\Delta}_{11} \leq 0.05, \quad \text{for sample size } n = 20 \quad ;$$

$$\hat{\Delta}_{11} \leq 0.023, \quad \text{for sample size } n = 50 \quad .$$

5. ENTROPY-DIFFERENCE INTERPRETATION OF VASICEK STATISTIC

To test the hypothesis H_0: X is $N(\mu,\sigma^2)$ Vasicek (1977) proposes a statistic which is equivalent to

$$\hat{\Delta}_{12} = \hat{H}_1^0 - \hat{H}_2$$

where $\hat{H}_1^0$ is an estimator of the parametric evaluation $H^0(f)$ of entropy, and

$\hat{H}_2$ is a fully non-parametric estimator of H(f) based on the gap or leap (of order 2ν) estimator

$$\tilde{q}_\nu(\frac{j}{n+1}) = \frac{n+1}{2\nu} \{X_{(j+\nu)} - X_{(j-\nu)}\} \quad , \quad j=\nu+1,\dots,n-\nu$$

of q(j/(n+1)), and

$$\hat{H}_2 = \frac{1}{n-2\nu} \sum_{j=\nu+1}^{n-\nu} \tilde{q}_\nu (\frac{j}{n+1})$$

Some significance levels of $\hat{\Delta}_{12}$ are given in Table 2; they are transformations of the significance levels given by Vasicek (1977) and obtained by Monte-Carlo simulation.

6. ENTROPY-DIFFERENCE INTERPRETATION OF PARZEN GOODNESS OF FIT PROCEDURE

To test the general hypothesis H_0: X is $F_0(\frac{x-\mu}{\sigma})$, Parzen (1979) proposes forming raw estimators $\tilde{d}(u)$ of

$$d(u) = \frac{1}{\sigma_0} f_0Q_0(u)\ q(u) \ ,$$

where $\sigma_0 = \int_0^1 f_0Q_0(t)\ q(t)\ dt$. To form $\tilde{d}(u)$ and $\tilde{\sigma}_0$ we replace q(u) by the least smooth gap estimator $\tilde{q}_2$ (u). Smooth estimators $d_m(u)$ of d(u) are formed by the autoregressive method. From estimators of the pseudo-correlations

$$\rho(v) = \int_0^1 e^{2\pi iuv}\ d(u)\ du, \quad v=0,\underline{+}1,\dots,\underline{+}m$$

one estimates the coefficients of the autoregressive order m approximator

$$d_m(u) = K_m \left| 1 + \alpha_m(1)\ e^{2\pi iu} + \dots + \alpha_m(m)\ e^{2\pi ium}\right|^{-2}$$

to d(u). The coefficient K_m plays an important role in entropy calculations since

$$\int_0^1 - \log d_m(u)\ du = -\log K_m$$

can be regarded as an estimator $\hat{\Delta}_{33} = \int_0^1 - \log \hat{d}(u)\ du$ of Δ.

This formula, which we prove below, provides an entropy-difference interpretation of the goodness of fit procedures in Parzen (1979).

To prove this interpretation of Δ_{33}, write

$$- \log d(u) = \log \sigma_0 - \log f_0Q_0(u) - \log q(u)$$

Therefore

$$\int_0^1 - \log d(u)\, du = H^0(f) - H(f)$$

is an entropy-difference.

The autoregressive estimator $\hat{d}_m(u)$ of $d(u)$ provides a parametric select estimator of $q(u)$ by

$$\hat{q}(u) = \tilde{\sigma}_0\, \hat{d}_m(u)\, q_0(u)$$

A parametric select estimator of $H(f)$ is

$$\hat{H}_3 = \int_0^1 \log \hat{q}(u)\, du$$

$$= \int_0^1 \log \hat{d}_m(u)\, du + \hat{H}_3^0$$

where

$$\hat{H}_3^0 = \log \tilde{\sigma}_0 + H(f_0)$$

is an estimator of $H^0(f)$ based on $\tilde{\sigma}_0$.

The parametric select entropy-difference test statistic $\hat{\Delta}_{33}$should be denoted $\hat{\Delta}_{33,m}$ because it depends on the order m of the autoregressive estimator $\hat{d}_m(u)$ of $d(u)$. Significance levels of $\hat{\Delta}_{33,m}$ derived by a very approximate Monte Carlo simulation (in the case of testing for normality) are given in Table 2. They show that the parametric select estimators of Δ provide a smooth progression of significance levels from the fully parametric estimators of Δ to the fully non-parametric estimators. In practice, we recommend adaptive determination of the order m by the data, rather than choosing a fixed order m.

It may be useful to use a rough approximation to the 5% significance levels of $\hat{\Delta}_{33,m}$ which is provided by $2m/n$. A criterion for accepting H_0: X is $F_0(\frac{x-\mu}{\sigma})$ is:

$$\hat{\Delta}_{33,m} = - \log \hat{K}_m \leq \frac{2m}{n} \quad , \; m=1,2,\ldots \quad .$$

One rejects H_0 if there exists a value of m for which the Akaike-type criterion

$$\text{AIC}(m) = \frac{2m}{n} + \log \hat{K}_m \leq 0 \qquad ;$$

the value of m which minimizes AIC(m) is chosen as an "optimal" value $\hat{m}$. An optimal parametric-select estimator of the true quantile-density function q(u) is

$$\hat{q}_{\hat{m}}(u) = \tilde{\sigma}_0 \, \hat{d}_{\hat{m}}(u) \, q_0(u) \quad .$$

7. CONCLUSION

We believe that the interpretation given in this paper of powerful goodness of fit procedures as entropy-difference statistics provides a striking demonstration of the FUN.STAT synthesis of statistical reasoning. In addition to elegance of the theory, very practical and implementable procedures are obtained.

The parametric select estimators $\hat{\Delta}_{33,m}$ of entropy-difference test statistics for goodness of fit have for m=1 approximately the properties of fully parametric estimators (such as Shapiro-Wilk $\hat{\Delta}_{11}$) and have for large values of m approximately the properties of fully non-parametric estimators (such as Vasicek $\hat{\Delta}_{12}$). Thus it appears the series $\hat{\Delta}_{33,m}$ provide all the test-statistics required. Further the autoregressive approach provides non-parametric estimators of the true distribution when one rejects the null hypothesis H_0.

One may find that a sample passes the goodness of fit procedure for two null hypotheses. An appealing procedure, whose properties remain to be investigated, is to choose that null hypothesis for which $\hat{\Delta}_{33,m}$ is always less than the corresponding statistic for the other hypothesis.

The entropy-difference statistics $\hat{\Delta}_{33,m}$ are implemented in our one-sample univariate data analysis computer program ONESAM. Table 3 lists autoregressive estimates of entropy-difference when testing for normality data sets in Stigler (1977). An asterisk indicates a data set which is not normal in our judgement.

In Table 2 we report significance levels for $\hat{\Delta}_{12}$ obtained (by Monte Carlo calculations) by Dudewicz and van der Muelen (1981) in the case of testing for uniformity rather than normality.

The closeness of the Dudewicz-van der Muelen levels to the Vasicek levels suggests a conjecture, which remains to be proved, that the entropy-difference statistics have distributions which are approximately the same for all null hypotheses H_0: X is $F_0(\frac{X-\mu}{\sigma})$.

A final noteworthy feature is that the autoregressive method of estimating quantile-density functions and density-quantile functions, introduced in Parzen (1979), can be shown to have a maximum entropy property [compare Parzen (1982)].

Table 2. 5% SIGNIFICANCE LEVELS FOR ENTROPY DIFFERENCE STATISTICS

Accept H_o: X is $N(\mu,\sigma^2)$ for some μ and σ if entropy difference is less than threshold given.

	$\hat{\Delta}_{11}$	$\hat{\Delta}_{33,m}$ Autoregressive order m Monte Carlo 5% level (rough approximation 2m/n)					$\hat{\Delta}_{12}$ Vasicek gap estimator $\tilde{q}_\nu(u)$ (Dudewicz-van der Muelen)				
Sample Size n	Shapiro-Wilk	m=1	m=2	m=3	m=4	m=5	ν=5	ν=4	ν=3	ν=2	ν=1
20	.05	.141 (.10)	.235 (.20)	.299 (.30)	.378 (.40)	.398 (.50)		.40 (.43	.40 .43	.43 .47	.61 .66)
50	.023	.045 (.04)	.081 (.08)	.126 (.12)	.153 (.16)	.176 (.20)	.21 (.22	.21 .22	.23 .24)		

Shapiro-Wilk and Vasicek levels are based on Monte Carlo simulation of normal; Dudewicz-van der Muelen levels are based on Monte Carlo simulation of uniform.

One can conjecture a relation between gap order 2ν and autoregressive order m for the corresponding estimators to have similar distributions and therefore similar significance levels:

$$(2\nu)\ m = n = \text{sample size}$$

To understand what this conjecture is alleging note that for n=20, m=4 is similar to 2ν = 6; for n=50, m=6 is similar to 2ν = 8.

When one uses gap estimators of q(u), and thus of entropy, one has the problem of determining the order 2ν. One can more easily develop criteria for determining the order m of autoregressive estimators of q(u).

Table 3. ANALYSIS OF STIGLER (1977) DATA SETS BY ONESAM PROGRAM

		$\lvert\tilde{\rho}(\nu)\rvert^2$			$\hat{\Delta}_{33,m}$			AIC(m)			Opt. Order
Stigler Data Set	Sample Size	$\nu=1$	$\nu=2$	$\nu=3$	m=1	m=2	m-3	m=1	m=2	m=3	$\hat{m}$
1	18	.042	.025	.057	.04	.08	.17	.07	.15	.17	0
*2	17	.193	.030	.042	.21	.27	.34	-.10	-.03	.02	1
3	18	.108	.027	.047	.11	.14	.17	-.00	.08	.16	0
4	21	.057	.159	.041	.06	.20	.21	.04	-.01	.08	2
5	21	.146	.015	.041	.16	.17	.22	-.06	.01	.07	1
6	21	.047	.102	.002	.05	.13	.15	.05	.06	.14	0
7	21	.041	.046	.040	.04	.11	.18	.05	.08	.11	0
8	21	.079	.047	.011	.08	.18	.27	.01	.01	.02	0
*9	20	.285	.235	.124	.34	.42	.42	-.24	-.22	-.12	1
10	20	.027	.059	.045	.03	.09	.15	.07	.11	.15	0
11	26	.046	.006	.033	.05	.06	.11	.03	.09	.12	0
12	20	.107	.001	.023	.11	.13	.13	-.01	.07	.17	1
13	20	.084	.027	.063	.09	.16	.20	.01	.04	.10	0
*14	20	.162	.094	.130	.18	.22	.39	-.08	-.02	-.09	3
15	20	.066	.006	.001	.07	.09	.09	.03	.11	.21	0
*16	20	.080	.056	.093	.08	.17	.44	.01	.03	-.14	3
17	23	.065	.014	.038	.07	.11	.14	.02	.07	.12	0
19	29	.002	.019	.008	.00	.02	.03	.07	.12	.18	0

REFERENCES

Dudwicz, E. J. and Van der Muelen, E. C. (1981). Entropy-Based Tests of Uniformity Journal of the American Statistical Association, 76, 967-974.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality, Technometrics, 17, 111-117.

Parzen, E. (1979). Nonparametric statistical data modeling. Journal of the American Statistical Association, 74, 105-131.

Parzen, E. (1982). Maximum entropy interpretation of autoregressive spectral densities. Statistics and Probability Letters, 1, 2-6.

Shapiro, S. S. and Francis, R. S. (1972). Approximate analysis of variance test for normality. J. American Statistical Association, 67, 215-216.

Shapiro, S. S. and Wilk, M. B. (1968). An analysis of variance test for normality, Biometrika, 52, 591-611.

Shapiro, S. S., Wilk, M. B. and Chen, H. J. (1968). A comparative study of various tests for normality, J. American Statistical Association, 63, 1343-1372.

Stigler, S. M. (1977) Do robust estimators work with real data, Annals of Statistics, 5, 1055-1098.

Vasicek, O. (1976). A Test for Normality Based on Sample Entropy, Journal of the Royal Statistical Society, B, 38, 54-59.

END

FILMED